Examples for reasons why molecules are not successfully extracted from patent documents

		format in patent document:	
		image	**text**
molecular structures described as:	**structural depiction**	Image-to-structure conversion • is not part of the workflow • fails due to bad image quality, low image resolution, or complexity of the structural depiction (e.g. Markush structures)	The structural ‘encoding’ that is being used is not recognised (e.g. 3-letter amino acid codes for peptides or SMILES)
	chemical name	Chemical names are not extracted from images. For example, tables containing chemical names might be images	The name-to-structure conversion can fail due to • typos in patents • OCR errors • parsing errors leading to partial names being used as input for the conversion (e.g. names for R-groups, part of a name due to a line break) • chemical names recognition doesn’t distinguish between R-groups and molecules (resulting in radicals, for example) • use of non-systematic names leads to failure of the name-to-structure conversion if the names are not in the dictionary